Clone FAQs

Contents

Clone nomenclature: How does NCBI manage genomic library and clone names?

Each genomic library is assigned a standard name and abbreviation that is unique for a given organism. Wherever possible, NCBI adopts library names and abbreviations that have been provided by library creators or distributors as the standard names.

A genomic clone's standardized name is comprised of its microtiter plate address (plate number, row and column), prefixed by the standard library abbreviation, to provide a unique clone name (Figure 1). If a non-standardized name has been provided in a clone's insert or end sequence record in an INSDC database, the non-standardized name is stored as an alias of the standardized name.

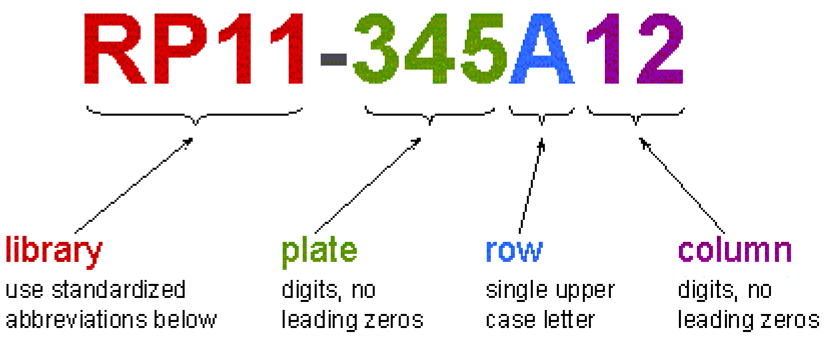


Figure 1 Standard clone name example

How do I find the clones discussed in 'Linking the human cytogenetic map with nucleotide sequence: the CCAP clone set' (PMID 16843097)?

Go to http://cgap.nci.nih.gov/Chromosomes/CCAP_BAC_Clones

Where can I find the clones that were positioned on the draft version of the human genome reference assembly, as described in the 2001 publication 'Integration of cytogenetic landmarks into the draft sequence of the human genome' (PMID: 11237021)?

The positioning of these clones involved the integration of cytogenetic, radiation hybrid, linkage and sequence data. They represent a historic resource used to help characterize genes associated with large chromosomal aberrations that result in human disease.

These clones are reported in a series of files in the Clone FTP site ftp://ftp.ncbi.nih.gov/repository/clone/reports/Homo_sapiens/; human_bac_resource_clone_rpt.txt, human_bac_resource_fish_map_pos.txt, human_bac_resource_sts.txt, human_bac_resource_summary_rpt.txt). These files provide the methods used to map each of these clone, as well as their reported cytogenetic locations, and the ids of the STS markers used for mapping. They also provide the chromosome assignments of these clones on the GRCh38 assembly.

How can I find information about the DNA source for a genomic clone library?

The %_libsource_$.xml file found in each of the organism-specific 'reports' directories in the Clone DB FTP site (ftp://ftp.ncbi.nih.gov/repository/clone/reports/) provides this information. Additional information on these files and other reports in this directory is provided in the following README: ftp://ftp.ncbi.nih.gov/repository/clone/reports/000_README_REPORTS.txt.

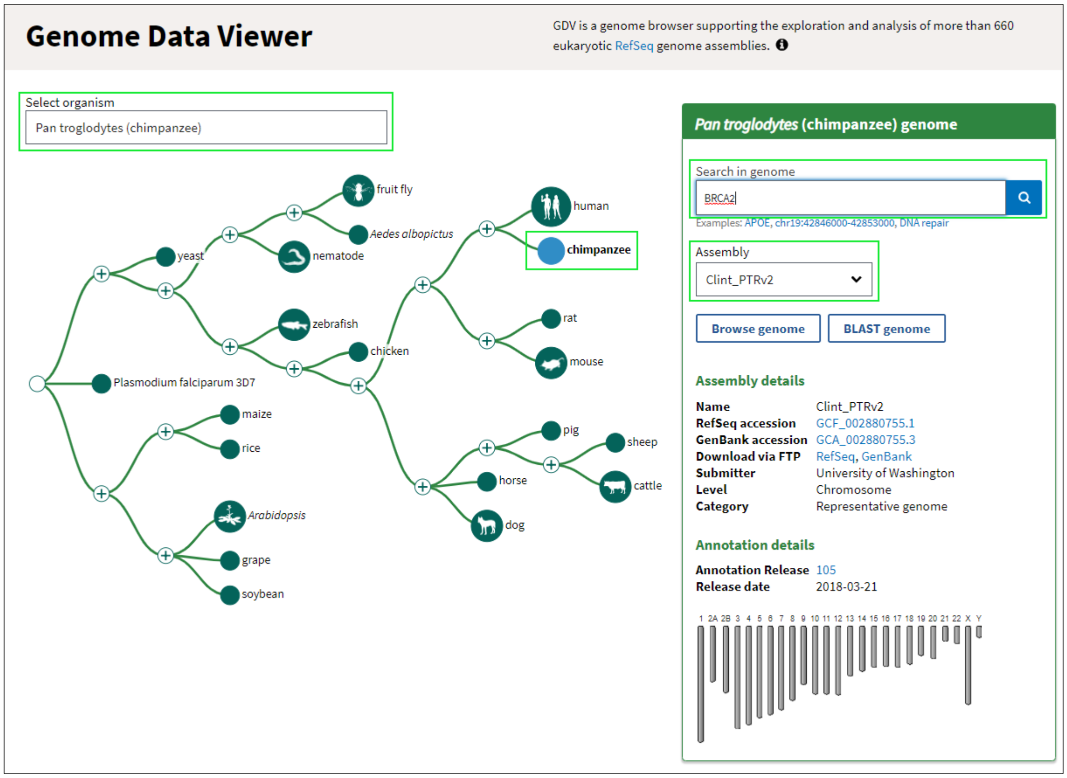
How can I find the insert or end sequences for genomic clones?

The '%.clone_acstate_$.out' and '%.endinfo_$.out' files found in each of the organism-specific 'reports' directories in the Clone FTP site (ftp://ftp.ncbi.nih.gov/repository/clone/reports/) provide the versioned accessions for insert and end sequences, respectively, for all clones in library % (for organisms with NCBI taxid $), mapped to the clone names. Additional information on these files and other reports in this directory is provided in the following README: ftp://ftp.ncbi.nih.gov/repository/clone/reports/000_README_REPORTS.txt.

To download the FASTA files corresponding to a list of end sequence accessions, provide these accessions as input to the endseq_dp.pl script in the Clone FTP 'utility' directory (ftp://ftp.ncbi.nih.gov/repository/clone/utility/). The README_UTILITIES.txt file in that directory provides detailed instructions for usage of this script.

How can I identify genomic clones that contain a gene or sequence of interest?

Such clones can be identified using either the clone placement tracks in the NCBI Genome Data Viewer (GDV) or the GFF placement files on the NCBI Clone FTP site. If using GDV, start at the home page (https://www.ncbi.nlm.nih.gov/genome/gdv/) to specify your organism and assembly of interest (Figure 2). Use the Search box at the top of this page to navigate to a gene or sequence location in the browser. For an introduction to GDV, check out the following NCBI YouTube playlist (http://bit.ly/GDV-tutorials).



Genome Data Viewer

GDV is a genome browser supporting the exploration and analysis of more than 660 eukaryotic RefSeq genome assemblies.

Select organism: Pan troglodytes (chimpanzee)

Pan troglodytes (chimpanzee) genome

Search in genome: BRCA2

Examples: APOE, chr19:42846000-42853000, DNA repair

Assembly: Clint_PTRv2

Assembly details

Name	Clint_PTRv2
RefSeq accession	GCF_002880755.1
GenBank accession	GCA_002880755.3
Download via FTP	RefSeq, GenBank
Submitter	University of Washington
Level	Chromosome
Category	Representative genome

Annotation details

Annotation Release	105
Release date	2018-03-21

Figure 2 Genome Data Viewer (GDV) home page

Once in the browser for your assembly of interest, use the 'Assembly Support' track set to add clone placement tracks to the display. To load this track set, click the 'Tracks' button in the upper right corner and under the 'NCBI Recommended Track Set' select 'Assembly Support' (Figure 3).

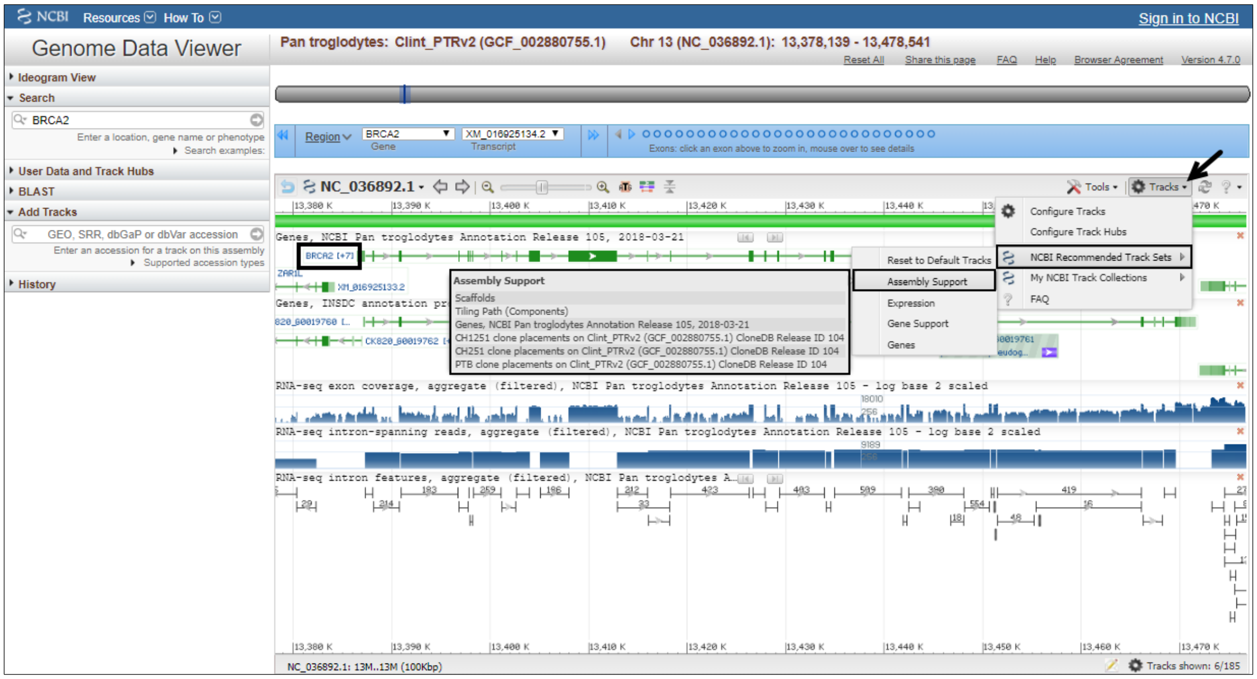

NCBI Resources How To
Sign in to NCBI
Genome Data Viewer
Pan troglodytes: Clint_PTRv2 (GCF_002880755.1) Chr 13 (NC_036892.1): 13,378,139 - 13,478,541
Reset All Share this page FAQ Help Browser Agreement Version 4.7.0
Ideogram View
Search
BRCA2
Enter a location, gene name or phenotype
Search examples:
User Data and Track Hubs
BLAST
Add Tracks
GEO, SRR, dbGaP or dbVar accession
Enter an accession for a track on this assembly
Supported accession types
History
Region BRCA2 Gene XM_016925134.2 Transcript
Exons: click an exon above to zoom in, mouse over to see details
NC_036892.1
Tools Tracks
Configure Tracks
Configure Track Hubs
Reset to Default Tracks
NCBI Recommended Track Sets
My NCBI Track Collections
FAQ
Assembly Support
Expression
Gene Support
Genes
Assembly Support
Scaffolds
Tiling Path (Components)
Genes, NCBI Pan troglodytes Annotation Release 105, 2018-03-21
CH1251 clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
CH251 clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
PTB clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
Genes, NCBI Pan troglodytes Annotation Release 105, 2018-03-21
BRCA2 (+7)
ZAR1L
XM_016925133.2
Genes, INSDC annotation pr
RNA-seq exon coverage, aggregate (filtered), NCBI Pan troglodytes Annotation Release 105 - log base 2 scaled
RNA-seq intron-spanning reads, aggregate (filtered), NCBI Pan troglodytes Annotation Release 105 - log base 2 scaled
RNA-seq intron features, aggregate (filtered), NCBI Pan troglodytes A...
13,380 K 13,390 K 13,400 K 13,410 K 13,420 K 13,430 K 13,440 K 13,450 K 13,460 K 13,470 K
NC_036892.1: 13M..13M (100Kbp)
Tracks shown: 6/185

Figure 3 Applying the 'Assembly Support' track set. The BRCA2 genome region is in view.

The ‘Assembly Support’ track set contains the placement tracks for the 3 clone libraries with the greatest number of placed clones on that assembly (Figure 4). If your clone of interest does not belong one of these 3 libraries, the placement tracks for additional libraries can also be accessed via the ‘Tracks’ menu. Instead of selecting ‘Assembly Support’ from this menu, choose ‘Configure Page’. In the resulting dialog, you’ll find the complete collection of clone placement tracks for that assembly listed in under the ‘Genomic Clones’ tab.

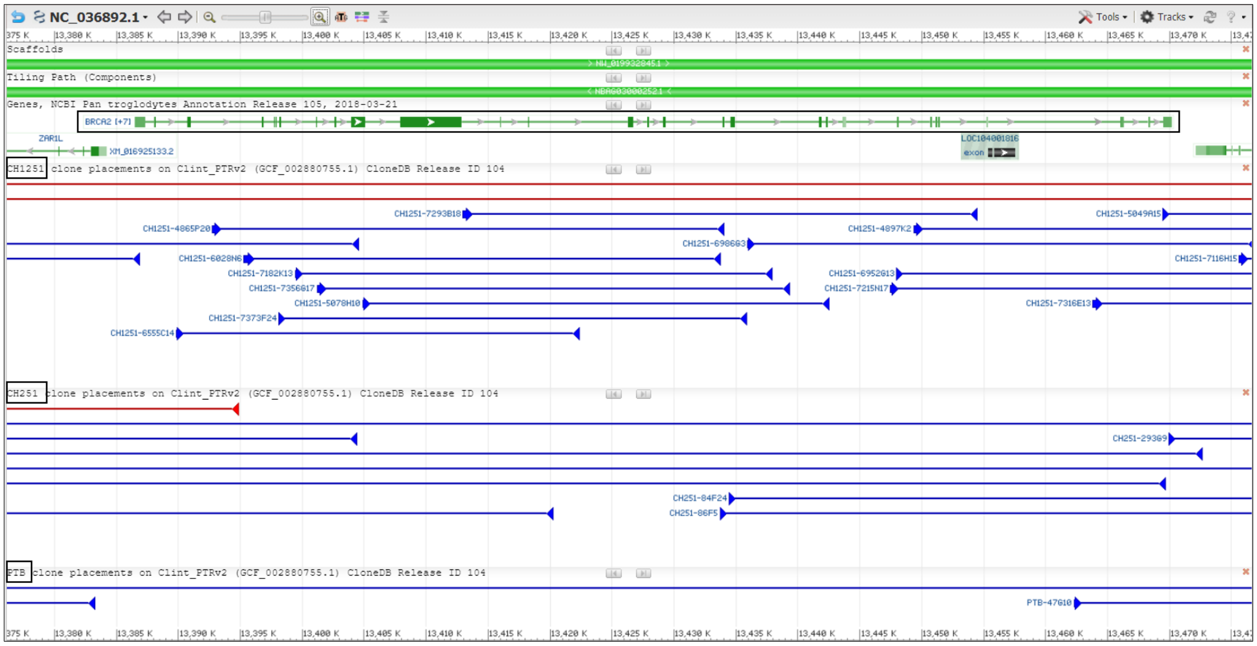

NC_036892.1
Tools Tracks
Scaffolds
Tiling Path (Components)
Genes, NCBI Pan troglodytes Annotation Release 105, 2018-03-21
BRCA2 (+7)
ZAR1L
XM_016925133.2
LOC104001816
CH1251 clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
CH1251-7293B18
CH1251-5049A15
CH1251-4865P20
CH1251-4897K2
CH1251-6986G3
CH1251-6028N6
CH1251-7116H15
CH1251-7182K13
CH1251-6952O13
CH1251-7356O17
CH1251-7215H17
CH1251-5078H10
CH1251-7316E13
CH1251-7373F24
CH1251-6555C14
CH251 clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
CH251-29J09
CH251-84F24
CH251-86F5
PTB clone placements on Clint_PTRv2 (GCF_002880755.1) CloneDB Release ID 104
PTB-47G10
13,380 K 13,385 K 13,390 K 13,395 K 13,400 K 13,405 K 13,410 K 13,415 K 13,420 K 13,425 K 13,430 K 13,435 K 13,440 K 13,445 K 13,450 K 13,455 K 13,460 K 13,465 K 13,470 K

Figure 4 Clone placement tracks shown in the BRCA2 genome region.

GFF and report files with clone placement data are also available from the ‘reports’ directories on the Clone FTP site (ftp://ftp.ncbi.nih.gov/repository/clone/reports/). The content of these directories and files is described in the following YouTube video (http://bit.ly/clone-ftp).

Additionally, within each organism directory, per-library summary data for clone and end placements are provided in the ‘clone_placement_report_$.out’ and ‘end_placement_report_$.out’ files. Additionally, a collection GFF files in these directories provides the placement details for individual clones. These GFF files are described in detail in the accompanying README file: ftp://ftp.ncbi.nih.gov/repository/clone/reports/000_README_GFF.txt

How do I interpret clone placement tracks?

The YouTube video http://bit.ly/clone-placements explains how to use clone placements to interpret genome assemblies for regions that contain putative structural variation or assembly errors.

The table below describes the color and shading scheme used in the clone placement tracks to depict various placement attributes and rendering examples of various attribute combinations.

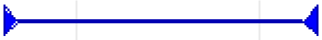
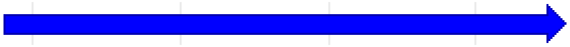
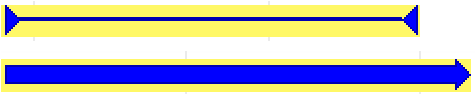
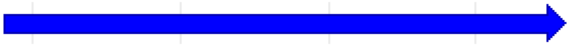
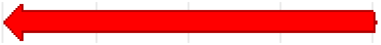
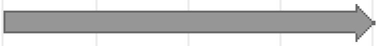
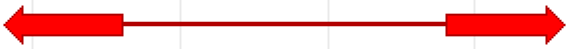
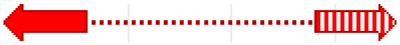
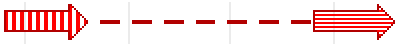
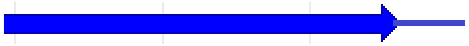
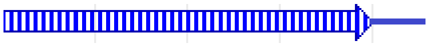
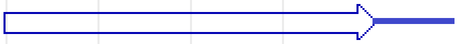
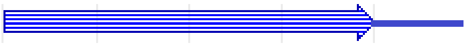
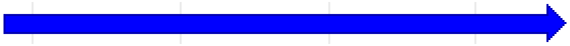
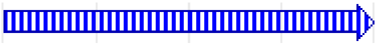
- **Real vs. "virtual" ends:** End-sequence derived clone placements are depicted by a line connecting two triangles whose orientations represent the orientations of the individual end placements. All concordant placements will have ends that face one another but ends comprising discordant placements may be found in various orientation combinations. Most end-sequence derived clone placements represent a pair of real end placements. However, a clone placement may occasionally be derived from a single end placement (e.g. a clone that hangs into an assembly gap). In such cases, the end without a placement is referred to as a "virtual" end.

- **Representing other placement methods:** Clone placements not derived from end placements (such as those derived from insert sequence placements) appear as rectangles with an arrow at one end to indicate orientation, if known.

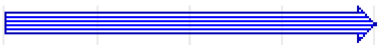
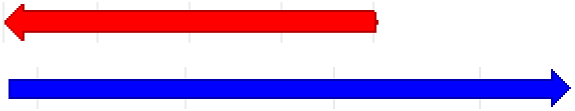
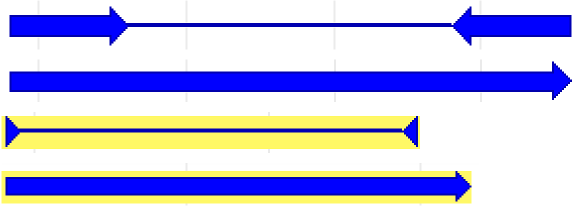
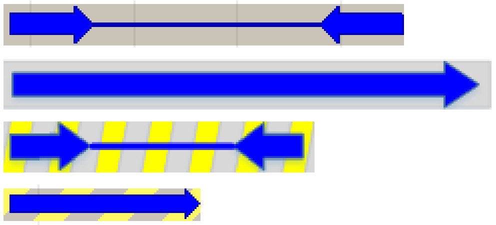
- **Supporting evidence:** The presence of non-supporting end and/or insert placements does not always indicate that a clone placement is incorrect. Such evidence may be due to the presence of some type of repetitive sequence in a clone sequence (resulting in multiple non-overlapping placements) or reflect the placements of unrelated end sequences that were incorrectly assigned a particular clone name during the sequencing process. It is recommended that users review all associated clone placement evidence when selecting a particular clone for their research.

Table 1 Clone placement display scheme

Clone Placement Display Scheme

Attribute	**Possible Values**	**Rendering**	**Visual Example**
Clone placement type	End-seq only	2 arrows joined by line	
	Insert only	Single arrow	
	Combined	2 arrows joined by line or single arrow on yellow background	
Concordancy	Concordant	Color: Blue	
	Discordant	Color: Red	
	Concordancy not set	Color: Gray	
End-seq only clone placement uniqueness	Unique	Connecting Line: solid	
	Multiple	Connecting Line: dotted	
	Uniqueness not set	Connecting line: dashed	
Clone end confidence	Unique	Fill: solid color	
	Multiple	Fill: vertical bars	
	Virtual	Fill: empty	
	Other/Not set	Fill: horizontal bars	
Insert only clone placement uniqueness	Unique	Fill: solid color	
	Multiple	Fill: vertical bars	

Clone Placement Display Scheme

Attribute	**Possible Values**	**Rendering**	**Visual Example**
	Other/not set	Fill: horizontal bars	
Directionality	Forward or Backward	Arrow	
Supporting Evidence	All non-prototype ends are ‘supporting’	With no shaded background	
	Not all non-prototype ends are ‘supporting’	With shaded background	

How does NCBI generate genomic clone placements?

NCBI produces three types of genomic clone placements, which are based on clone ends, clone inserts, or a combination of the two.

- **End clone placements:** clone end sequences retrieved from GenBank (Genome Survey Sequences (GSS)) and Trace Archives records are screened by Clone DB to remove low quality bases and vector contamination. Only end sequences with ≥ 50 cleaned bases are considered suitable for further analysis. Cleaned sequences are aligned against a window-masked (https://www.ncbi.nlm.nih.gov/sites/pubmed/16287941) copy of the assembly of interest using NG Aligner, an NCBI-developed, BLAST-derived alignment algorithm. End alignments are performed at the assembly scaffold level and then mapped to the top-level molecule to which the scaffold belongs (i.e. chromosome, in the case of placed scaffolds). End placements are ranked by weighted identity, a combined measure of the end alignment’s percent identity and coverage, on a per-assembly basis. Ties are allowed when an end has more than one placement with the same weighted identity in an assembly.
  Clone placements are generated by the pairing of end placements representing the forward (F) and reverse (R) ends of a given clone. Any pair of ranked end placements located on the same top-level molecule may be considered for use in the generation of a clone placement. However, to avoid the creation of duplicate clone placements in instances where there are multiple

sequences representing the same clone end, NCBI clusters overlapping ranked end placements and selects a single prototype for usage in clone placements. An end cluster is defined as a set of similarly oriented, ranked end placements from one or more end sequences representing the same end (F/R) of a single clone that overlap in part or in whole on a given assembly scaffold. The cluster prototype is the end placement that holds the 5'-outermost position on the scaffold to which it aligns and is the only placement from a cluster used for clone placements.

A clone placement is comprised of single F and R end placements (Figure 5; red and green arrows connected by line). Within an end cluster (3 green "R" arrows), the prototype end placement (outlined in black), holds the outermost position and will contribute to the longest possible clone placement.

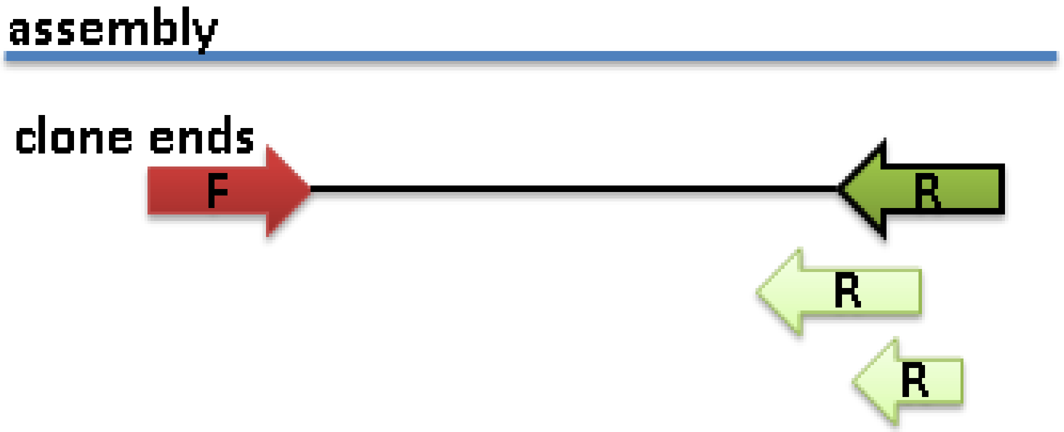


Figure 5 Example clone placement and end cluster

NCBI reports the *most likely* placement(s) for each clone. If prototype end placements contribute to multiple self-overlapping clone placements, NCBI selects a single clone placement from the overlapping set as an archetypal placement. Among clones with concordant clone placements, the archetype represents the concordant clone placement comprised of the best ranked pair of end placements in the set. If no concordant placement exists, the archetype represents the discordant clone placement comprised of the best ranked pair of end placements. If more than one clone placement in the set is comprised of the best ranked pair of end placements, the shortest such placement comprised of correctly oriented ends will be selected as the archetype. A clone may have more than one archetypal clone placement, but they may not overlap one another. Only archetypal placements are reported and displayed in clone records.

- **Insert clone placements:** Clone insert sequences are retrieved from GenBank. Only inserts representing the latest version of HTG phase 3 (finished) sequences that are associated with a single clone id are considered for clone placements. The insert sequences are aligned against a window-masked copy of the assembly of interest using NG Aligner, an NCBI-developed, BLAST-derived alignment algorithm. Insert alignments are performed at the assembly scaffold level and then mapped to the top-level molecule to which the scaffold belongs (i.e. chromosome, in case

of placed scaffolds).
Insert placements are ranked by weighted identity, a combined measure of the insert alignment's percent identity and the coverage, on a per-assembly basis. Placements for insert sequences that are a component of the assembly unit on which the clone is being placed are defined by the alignment with greatest overlap with the coordinates of the component in the assembly unit. For insert sequences that are not a component of the assembly unit on which the clone is being placed, NCBI clusters any overlapping ranked insert placements. An insert cluster is defined as a set of similarly oriented, ranked insert placements from the same clone id that overlap in part or in whole on a given assembly scaffold. The alignment with the greatest weighted identity in each cluster is used to define the archetypal insert placement. Only archetypal placements are reported and displayed in clone records.

- **Combined clone placements:** These placements are provided only for clones that have both end and insert sequences available. End and insert placements are generated independently as described in the previous sections, and then one or more is defined as the combined placement for the clone, according to the following rules:
    - If the insert-based archetypal placement is overlapping and fully contained within the archetypal end-based clone placement:
        1. If the end-based placement is concordant, the end-based placement shall be selected as the combined clone placement, regardless of the concordancy of insert placement.
        2. If the end-based placement is discordant and >= 500 kb (BACs, PACs) or 100 kb (fosmids), the insert-based placement shall be selected as the combined clone placement, regardless of the concordancy of the insert placement.
        3. If neither criterion applies, the end-based placement shall be selected as the clone placement.
    - If the archetypal end-based placement is overlapping and entirely contained within the insert-based placement (Insert>End), the insert-based placement shall be selected as the combined clone placement, regardless of concordancy status of either placement.
    - If the archetypal end-based clone placement and insert-based clone placement are partially overlapping (a.k.a "skewed placements"), the following criteria shall be used to define the combined clone placement:
        1. If the end-based clone placement is concordant and the insert-based clone placement is discordant, the end-based clone placement shall be declared the combined clone placement (regardless of the reason for the discordancy).
        2. If the end-based clone placement is discordant and the insert-based clone placement is concordant, the insert-based placement shall be declared the combined clone placement (regardless of the reason for the discordancy).

3. If both the end-based clone placement and the insert-based clone placement are discordant, the insert placement shall be declared the combined clone placement (regardless of the reason for either discordancy).
4. If both the end-based clone placement and the insert-based clone placement are concordant, the insert-based placement shall be declared the combined clone placement.

- If the archetypal end-based clone placement and archetypal insert-based clone placement do not overlap, the clone shall be declared as "multiply placed”.
- If there are multiple archetypal insert-based placements that overlap each other on opposite strands and that also overlap an end-based archetypal clone placement (<= 500 kb (BACs, PACs) or <= 100 kb (fosmids)), all such placements will be retained as clone placements.
- If there are multiple non-overlapping archetypal insert placements that overlap the same archetypal end-based clone placement (<= 500 kb (BACs, PACs) or <= 100 kb (fosmids)), all such placements will be retained as clone placements.
- If there is a combination of the two preceding conditions, all such placements will be retained as clone placements.

Why doesn't my clone have a placement?

Clone placement involves two steps: sequence alignments and placements. A clone may lack a placement if it has issues at either step. Only high-quality clone end sequences (≥ 50 cleaned bases) and phase 3 (finished) clone insert sequences are considered suitable for clone placements. Once alignments are produced, clone placements are generated from ranked alignments. Rank is a combined measure of alignment percent identity and percent coverage. Common reasons for the absence of a clone placement include: (1) lack of high quality sequence for one or both clone ends, (2) lack of finished insert sequence, (3) no alignment of one or both ends or insert sequence to assembly or (4) quality of end or insert alignment is too low for use in placement step. Note: (3) and (4) are more likely to be observed when clone sequences fall within repetitive genomic regions. Clones for which one or both ends were placed, but no clone placement was created, are reported in files on the FTP site. See ftp://ftp.ncbi.nih.gov/repository/clone/reports/000_README_GFF.txt for details.

How is the average insert length and standard deviation for a specific genomic library defined?

The average insert length and standard deviation for each genomic library are calculated from a subset of the clone placements generated on the reference assembly. Only the following clone placements are used for this calculation:

- both ends have a unique placement on the primary assembly unit
- both ends are placed on the same scaffold
- both ends are correctly oriented (i.e. face each other)

- for BACs/PACs, clone size is >50 kb and <500 kb
- for fosmids, clone size is >10 kb and <100 kb.

Note: Because the library average insert and standard deviation are defined by clone placements, these values reflect the assembly to which the library was aligned and may change with assembly updates.

How is placement concordance defined?

Genomic clone placements created by NCBI are defined as either concordant or discordant. Concordant clone placements meet the following criteria:

- End placements are found in opposite orientation and facing one another.
- End placements are located within 3 s.d. of the library insert average from one another.
- Insert placement length is within 3 s.d. of the library insert average (as defined by end-sequence placements).